Transparency, Usability, and Reproducibility: A Four-Step Plan toward Improved Comparative Databases Using Primates as Examples

Carola Borries[*], Aaron A. Sandel[†], Andreas Koenig[*], Eduardo Fernandez-Duque[‡], Jason M. Kamilar[§], Caroline R. Amoroso[¶], Robert A. Barton[#], Joel Bray[**], Anthony Di Fiore[††], Ian C. Gilby[**, ‡‡], Adam D. Gordon[§§], Roger Mundry[¶¶], Markus Port[##], Lauren E. Powell[#], Anne E. Pusey[¶], Amanda Spriggs[§§], Charles L. Nunn[¶, ***].

[*] Department of Anthropology and Interdepartmental Doctoral Program in Anthropological Sciences, Stony Brook University, SUNY, Stony Brook, NY 11794-4364
[†] Department of Anthropology, University of Michigan, Ann Arbor, MI 48109
[‡] Department of Anthropology, Yale University, New Haven, CT 06511; Facultad de Recursos Naturales, Universidad Nacional de Formosa, Formosa 3600, Argentina
[§] Department of Anthropology and Graduate Program in Organismic and Evolutionary Biology, University of Massachusetts, Amherst, MA 01003
[¶] Department of Evolutionary Anthropology, Duke University, Durham, NC 27708
[#] Evolutionary Anthropology Research Group, Durham University, Durham, DH1 3LE, UK
[**] School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85281
[††] Department of Anthropology, University of Texas, Austin, TX 78712
[‡‡] Institute of Human Origins, Arizona State University, Tempe, AZ 85281
[§§] Department of Anthropology, University at Albany, SUNY, Albany, NY 12222
[¶¶] Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany
[##] Department of Behavioral Ecology, University of Goettingen, Goettingen 37077, Germany
[***] Duke Global Health Institute, Duke University, Durham, NC 27708

Address for correspondence: Carola Borries, Department of Anthropology, S-515 SBS Building, Circle Rd, Stony Brook University, Stony Brook, NY 11794-4364
email: carola.borries@stonybrook.edu, phone: (+) 631 632-1518, fax: (+) 631 632-9165

Total count: 14 text pages

Key words: metadata; primary sources; procedure documentation; user interaction

Short title: Four steps to improve comparative databases

Recent decades have seen rapid development of new analytical methods to investigate patterns of interspecific variation. Yet these cutting-edge statistical analyses often rely on data of questionable origin, varying accuracy, and weak comparability which seems to have reduced reproducibility of studies. It is time to improve the transparency of comparative data, while also making these improved data more widely available. We, the authors, met to discuss how transparency, usability, and reproducibility of comparative data can best be achieved. We propose four steps: (1) data identification with explicit, operational definitions and complete descriptions of the methods; (2) inclusion of fields to capture key characteristics of the data, such as sample size or nutrient availability (e.g. captive *versus* wild animals); (3) documentation of the original reference for each datum; and (4) facilitation of effective interactions with the data via user friendly and transparent user interfaces. We urge reviewers, editors, publishers, database developers and users, funding agencies, researchers publishing their primary data, and those performing comparative analyses to embrace these standards to increase the transparency, usability, and reproducibility of comparative studies.

From the beginning of evolutionary biology, the comparative method has been a major analytical tool,[1-3] allowing for the examination of patterns and processes of evolutionary change.[4] Some of the main obstacles to overcome in comparative analyses have been statistical in nature: How should we control for confounding variables? What criteria should we use to assess whether patterns are statistically significant and biologically meaningful? How should we control for the non-independence of comparative data that stems from phylogenetic relatedness? Much progress has been made with respect to these issues, especially in the development and use of

phylogenetic comparative methods.[3,5,6] For example, building on initial descriptions of phylogenetically independent contrasts,[7] methods can now incorporate phylogenetic uncertainty,[8,9] intraspecific variation,[10-12] and different models of phenotypic evolution.[13,14]

Although phylogenetic and statistical methods are rapidly advancing, an increasing number of researchers argue that the data to which these methods are applied are 'stuck in the dark ages'.[15-17] It is imperative that, before the specific methods employed in a comparative study are considered, the suitability of the data be thoroughly evaluated. The time has come to bring our comparative databases into the modern age, and to represent uncertainty in the data in the same way we might represent uncertainty in a statistical model or in a phylogeny.[18] It is also important that we be able to evaluate which sources of uncertainty – in the data, the phylogeny, and the statistical methods – have the greatest influence on comparative results.

To approach these issues, the authors met on May 28, 2014 at the National Evolutionary Synthesis Center (NESCent, Durham, NC, USA) and developed a four-step plan to improve comparative databases. We focused on primates, a relatively well-studied mammalian order that is the subject of many comparative studies. However, these concerns and suggestions are relevant to all taxonomic groups and disciplines.[19-21] Here, we begin by identifying the problems that are often shared by investigators of a wide range of comparative questions involving morphology, life history, behavior, and ecology.

One problem is that data points cannot always be traced to actual measurements. Gestation length in proboscis monkeys (*Nasalis larvatus*) appears to be such a case. As far as we know, the gestation length for this species has yet to be determined. Nevertheless, in almost all primate life history compilations, gestation length for proboscis monkeys is reported as 166 days (e.g.[22-24]), and this value has been used in many comparative studies (e.g.[25,26]). The value of 166 days

appears to originate from Schultz,[27: 281] who stated 74 years ago that: "Nothing is known in regard to the duration of the various periods of growth in the proboscis monkey, but it may be assumed that these do not differ radically from the conditions in macaques. In the latter pregnancy is known to last 166 days …". Such a statement was acceptable at the time, when the strong allometric relationship between body mass and life history traits was less widely appreciated. Today, however, it is unreasonable to assume a similar gestation length in two species of such different adult female body mass (10.5 kg in the proboscis monkey *versus* 4.9 kg in rhesus[28]). Such erroneous claims may be perpetuated in any study, but comparative studies are particularly vulnerable, as the authors are unlikely to have in-depth knowledge of every taxon included in the analysis (see also[29]). Unfortunately, such inappropriate attribution of data to particular sources is a recurring problem in comparative databases of primate life history (Borries et al., unpublished compilation).

In other cases, data from the primary literature that percolate into comparative studies may reflect results that are of questionable value because of small sample sizes, short study periods, or a specific research design. For example, a group of wild long-tailed macaques (*Macaca fascicularis*) was characterized as 100% fruit-eating during five months of the year based on 2, 3, 7, 7, and 20 instances of feeding observed per month, respectively; the same group was considered 100% grass-eating in another month based on a single feeding observation[30: 227]. Clearly, one cannot be confident in the diet of a population when assessed with only a handful of isolated feeding observations, yet these values have subsequently been used in comparative tests (e.g.[31,32]).

Similarly, certain data recording methods may lead to results of limited general use, for example when sizes, compositions, or densities are estimated for unhabituated groups (where

many of the group members may have already fled from the observer's approach) or during brief encounters on transects when many individuals are out of sight for similar reasons (e.g.[33]). Such approaches may result in group sizes being underestimated, and group densities being overestimated by a factor of 2 or more.[34]

All the above issues relate to the accuracy of individual data points; how closely do data match the "true" value which is free from systematic errors[35]. While it may be difficult to demonstrate the extent of bias introduced by a single inaccurate data point, the basic problem runs deeper. As scientists we are obliged to provide and use accurate data, because inaccurate data have the potential to reduce reproducibility leading to poor use of time and resources.[36] Importantly, in the few cases currently published in our field, existing databases were found to contain multiple data points with problems like those described above.[15,33] Using such inaccurate data in comparative analyses may bias the results, or lead to failure to detect existing patterns.

To evaluate the effect of accuracy, Borries et al.[15] compared gestation lengths in two primate taxa (Asian colobines and Asian macaques) with data drawn from four published life history compilations. Gestation length is expected to be similar among closely related species and vary with body size.[37] However, the authors found no significant relationship between gestation length and body size or taxon in any of the four datasets (using a phylogenetic generalized least squares model, all *P's*>0.05). In contrast, the model based on a fifth set of data containing only entries checked for accuracy produced the expected relationship (R^2_{adj}=0.91, P<0.001) with significant effects for body mass (*P*<0.007) and taxon (*P*<0.001).[15]

Comparative studies may also be compromised when data collected under different criteria or with incompatible methods are pooled under the same trait. Even a trait as seemingly straightforward as body mass may generate substantial errors (beyond resolution and precision of

the scales used[38]) when data for different definitions of what counts as an 'adult' individual are lumped.[39] In a study of ape morphology, 28% of chimpanzee skeletons and 38% of gorilla skeletons had adult dentition, but their bones were still growing and they had likely not yet reached adult mass.[40] Thus, despite not being fully grown, these individuals would have been classified as adults if dentition were used as the defining characteristic for adulthood. Another example of multiple definitions of a trait is 'weaning age', which in primates can include the following: the age at first intake of solid food (within the first weeks or months of life); observations of conflict over access to the nipple; the ability to survive as an orphan; or the age at cessation of nipple contact,[41,42] which in extreme cases can average 6.5 years (Bornean orangutan, *Pongo pygmaeus wurmbii*[43]).

Subsuming data based on such vastly different definitions into a single trait ('weaning age') will be unlikely (in most cases) to produce a false positive (Type I error). Compared to the earlier example on accuracy in gestation length, here now each data point can be very accurate, but compatibility may be compromised. Combining incompatible data as in the case of weaning age may prevent us from detecting real patterns[44] or from determining the relative strength of different effects.[45]

Another example, from the study on life history in Asian colobines and Asian macaques[15] illustrates this. Data for age at first reproduction and reproductive rates were drawn from only primary sources and checked for accuracy, trait definitions, and data collection methods. Still, the comparison revealed no statistically significant relationship between body mass and either of these life history variables. This could be explained by the fact that data from different nutritional conditions (captive and wild) were combined in the dataset, as nutritional intake is known to greatly impact maturation and reproductive output.[46] Including additional information

on nutrient availability as a simple binary variable resulted in greatly improved models with an $R^2_{adj.}$ value of 0.42 for age at first reproduction (*P* value for body mass =0.01) and an $R^2_{adj.}$ value of 0.65 for reproductive rates (*P* value for body mass <0.01).

Together, these issues related to accuracy and comparability may lead to conflicting results that cause more confusion than clarity and thus slow our progress toward finding general patterns. Consequently, reproducibility may become impossible and results from different studies may differ, resulting in less, rather than more certainty in our conclusions. While it is often unclear why results vary among comparative studies, we suspect that many differences emerge because researchers use different data collection protocols, employ divergent definitions for traits of interest, and rely on inaccurate or incomplete or imprecise compilations as datasets – all circumstances that can and should be improved.

FOUR STEPS TO IMPROVE COMPARATIVE DATA

To tackle these issues, we met and discussed the status of comparative databases for the mammalian order Primates. Below, we outline the four steps that we believe have the potential to improve future comparative studies. We were guided by our experience with various kinds of datasets, analyses, and questions, and by Whitlock's[47: 62] advice: "The central goal to have in mind … is to ensure that a new user, perhaps someone unknown to you working with the data 20 years later, can correctly interpret the results and derive correct conclusions from the data." Thus, in addition to ensuring the transparency of a database and its usability in the present, we were also concerned about future reproducibility of the result. This 'call to arms' agrees with several of the standards of the Transparency and Openness Promotion Guidelines,[48,49] which among others call for a standardization of research procedures and a clear description of all

aspects of data collection and definition.

Step 1): Identifying Data. The path to improvement begins with unequivocal, complete descriptions and definitions of all variables included in a comparative database, with details about how data were measured or determined.[50] Before naming a variable, it is recommended to check the literature to identify and use previously published definitions. Using precise operational definitions also provides explicit criteria for including and excluding data from a comparative database. Importantly, once a definition has been chosen, only data matching it should be included in the database. We suggest reporting in comparative databases the means, standard deviations, medians, and ranges when available, or calculations of these measures when the original source provides raw data. To guarantee transparency and reproducibility, it must be possible to trace each data point back to its original source. To this end, every alteration, even a mere conversion of dimensions (e.g., from days to weeks or centimeters to millimeters), needs to be identified in the database, for example by using a Boolean data type to indicate whether a certain action was performed or not.

When compiling data, the rate of transcription errors can be much reduced by using a relational database,[3] or a 'not only SQL' approach.[51] An additional, essential aspect of generating high quality databases is the proof-reading and double-checking process.[52] Ideally, someone other than the person who entered the data would perform the double-checking. As a final step, we recommend having an external expert examine the selected data. This could be an author of the underlying primary sources, or somebody familiar with the taxon and its relevant literature.

Step 2): Including Metadata. It is essential to include additional information (metadata) to

further characterize and contextualize the primary data used in a comparative analysis. Beyond those categories summarized in the Dublin Core Metadata Initiative (http://dublincore.org/) and Darwin Core (http://rs.tdwg.org/dwc/),[53] we strongly recommend including location of sampling (geographical coordinates and their precision), time period, study duration, number of groups, number of individuals, and other measures of sample size. Some of these types of metadata are being included in proposed extensions to Dublin and Darwin Core, including PaleoCore (http://paleocore.org/) for paleobiology datasets and EthoCore (http://ethoinformatics.org/) for behavioral and ecological datasets. Metadata are essential components of comparative databases to capture trait variation within and among species (see e.g.[50]). They furthermore allow for a gross quality assessment (e.g. sample size, number of individuals), and enable users of the data to select particular types, such as only those studied for a specified minimum time period, for specific analyses.

Some metadata also help to categorize the core methods used for data collection (e.g. gestation length based on conceptions estimated via hormonal concentrations *versus* based on mating patterns), or data analysis (e.g. home range sizes calculated using minimum convex polygons *versus* local convex hull *versus* kernel density methods). Information on ecological context and nutrient availability is also crucial, given that captive and wild animals have different nutritional regimes that may affect key variables, such as body mass or speed of growth and reproduction.[54-56] Although a distinction between captivity, provisioning, and food-enhanced conditions (crop raiding) is often possible,[39] it may suffice to distinguish if the study animals consumed any kind of human-made food.[15] This enables the compiler of comparative data to include, for example, data from captivity and the wild into a single database, and then control for nutritional conditions in the analysis.

Step 3): Documenting Procedures. To maintain reproducibility, a comparative database requires a written protocol that describes the specific search strategy used to locate data that were subsequently selected and included. Such documentation will also include the list of terms used in online search engines, how primary sources were located, and which other search variables or methods were employed, such as searches within a given species, by study site, or by variable. These protocols should be clearly written and linked to the database and/or provided as publications that describe or use the database.

In comparative databases, every datum is ideally documented by providing its source (the full reference for its first publication) together with the page number and/or Table/Figure numbers, as appropriate. This ensures that the primary source indeed exists (unlike the proboscis monkey example above) and allows for speedy location of the data even within extensive sources such as books or theses. Furthermore, it enables users to reconcile discrepancies in existing datasets.

Trait definitions, sampling methods, and actual data values can only be extracted directly from the primary source. On occasion, relevant metadata may have to be retrieved from other primary sources, and it is important that the trail to those sources also be provided. Past compilations can be helpful in locating primary sources, but they themselves cannot serve as primary data sources. The only exceptions are databases assembled in accordance with the guidelines outlined here. When an analysis is published based on a comparative database, the version number of the database should be identified explicitly.[52] Before using such a database, however, we recommend conducting multiple, random spot checks against the primary sources provided for a taxon that the author is very familiar with, and to only use databases with very

low error rates.

Step 4): Facilitating Effective Interactions with the Data. Ideally, comparative databases are made publicly available, leaving it to the user to decide if and which of its components are considered important. To facilitate access to a database by users who are unfamiliar with its basic contents and structure, all key components of the comparative database including the schema (which illustrates the relationships between the different components) can be summarized in a concise yet complete 'read me' file. In addition, an index and a table of contents will facilitate orientation and provide a first overview.

All elements of a database (including metadata) are best be made available for download in a widely available format, such as text files, thus allowing for a wide range of future uses via different programs. We recommend providing a clear, largely self-explanatory output design, with an easy to understand web browser based Graphical User Interface (GUI) that allows for a limited set of query options. The interface will also help prevent accidental misuse of the database, such that fields can only be combined in ways intended, and users will be unable to gain access to the underlying database directly (where errors could be introduced). A GUI can best be improved by running extensive test queries prior to the release that simulate data extractions required for already published analyses. These processes usually take more time than anticipated, but they are extremely important as their outcome may decide the success or failure of a database.

ADDITIONAL MEASURES TO IMPROVE COMPARABILITY

Implementing the steps outlined above requires the support of authors and reviewers of the

primary literature, editors and publishers, developers and users of comparative databases, and funding agencies. We now consider each of these in turn.

Authors publishing primary data that might later be collated into a comparative database play a key role in setting standards for the data available. Steps 1 and 2 above in particular demonstrate the importance of clear variable definitions and associated contextual information that can be used later as metadata. Authors are encouraged to publish this information, even if unnecessary for their current manuscript. The extra information can be presented as supplemental material and by referencing published work containing these data. In the near future, we hope that we as a discipline will agree on explicit guidelines to standardize the data even before they are being collected.

Reviewers are also essential to improving the primary and comparative data reported. As manuscript reviewers, we can all contribute toward implementing new rigor by requesting additional explanatory information from our peers to meet the standards above. Compliance may be better achieved when reviewers articulate to authors why following standards will increase the impact of their research, rather than simply setting a bar for authors to reach to achieve publication.

Editors and publishers can support the process by allowing for the inclusion of additional information. They can furthermore attach contingencies to acceptance of primary and comparative research papers, such as data upload in respective databases before a publication can go online. A good example is in place for DNA sequences, which are submitted to and made available through NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/). This is also an elegant solution to keep existing databases updated in the long run. Recording and publishing data so that they can be used in comparative databases is an important core contribution and

should be rewarded with recognition in the form of citations. This would require some changes to the data reporting process in comparative publications.[57] We support referencing the authors of the data compilations used, as well as all primary sources considered in the respective analysis. Change is already underway in several scientific fields where multiple journals have adopted a catalogue of increasingly stringent standards developed by the Transparency and Openness Promotion Committee.[48]

Database developers can also help achieve these new standards by working closely with scientific experts from the targeted research areas. Standardization can be facilitated by providing access to the underlying metadata of a database; in this way, authors of primary data become aware of what to report and which metadata to include in their work. In addition, developers can play a major role in making databases more comprehensive while at the same time easy to use.

Database users can play a major part by providing feedback on individual entries in the databases they use. Their input can be facilitated by web based portals allowing for flexible comments and information transfer to complete or correct specific database content.

Improving transparency and comparability is a slow process requiring recognition by research sponsors. We suggest that funding agencies could approach this in several ways. Standardization of data reporting and sharing could be implemented in data management plans. In addition, by including these standards in calls for proposals – and specifically calling for development of comparative databases – there will be incentives to invest in the steps outlined above. Funding agencies can also provide options for long-term maintenance and continued improvement of existing databases. Finally, any effort to improve transparency, usability, and reproducibility of data should be honored in decisions made by funding agencies, and by tenure

and promotion committees.

MOVING FORWARD

Currently, it is the collective responsibility of all researchers building and using comparative datasets to assess the impact of data accuracy and compatibility on the results of comparative studies. We already have evidence that data for wild animals differ from those for captive or provisioned ones,[56,58-60] that body mass data are prone to large errors,[39,40] and that estimates of group size are very sensitive to sampling methods.[33] We are also gaining a better understanding of the consequences of intraspecific variation on some key associations such as between neocortex size and group size.[61] We need more of such studies. Unfortunately, achieving similar results based on different datasets[25] is no guarantee for accuracy, as existing compilations are often strongly interdependent and may contain similar or even identical flaws.[15]

We are aware that the suggested steps toward transparency, usability, and reproducibility come at a price: the process requires a major time investment that will slow down comparative research until databases become available that are in compliance with the standards proposed here. We are also aware that more transparency of research methods and materials in comparative databases is just a first step. Different studies often use different methods (data collection, analysis), the results of which may be difficult or even impossible to compare. Reaching standardization at the level of data collection, as well as for analyses is an additional, important goal for the future. We are reminded of Felsenstein's[7: 14] original call to arms to use phylogenies in comparative studies, when he noted that, "Some reviewers of this paper felt that the message was 'rather nihilistic,' ..." Yet in the past 30 years, a huge diversity of new methods – and phylogenies – has emerged to fill the gap he identified. We are now at a similar point with

regard to improving comparative databases, so that these methods can be applied to their best effect and the findings are more certain. We urge authors, reviewers, editors, publishers, database developers, and users – as well as funding agencies and compilers of data – to embrace these standards and to honor the accompanying efforts, to help us generate new knowledge.

ACKNOWLEDGEMENTS

We gratefully acknowledge the National Evolutionary Synthesis Center (NESCent, Durham, USA), which is supported by the NSF (EF-0905606), for hosting and supporting our meeting on May 28, 2014. Additional support was provided by NSF Biological Anthropology (BCS-1355902) and Duke University. We thank Robyn Overstreet for fruitful discussions and contributions. We are also very grateful to Charles Janson, an anonymous reviewer, and John Fleagle, for very helpful comments and suggestions on the manuscript.

REFERENCES

1 Tinbergen N. 1963. On aims and methods of Ethology. Z Tierpsychol 20:410-433.

2 Darwin C. 1859. On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life. London: Murray. 502 p.

3 Nunn CL. 2011. The comparative approach in Evolutionary Anthropolgy and Biology. Chicago: University of Chicago Press.

4 Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. New York: Oxford University Press. 239 p.

5 Garamszegi LZ, editor. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology - concepts and practice. Berlin: Springer.

6 Garland TJ, Bennett AF, Rezende EL. 2005. Phylogenetic approaches in comparative physiology. J Exp Biol 208:3015-3035.

7 Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat 125:1-15.

8 Pagel MD, Lutzoni F. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. In: Laessig M, Valleriani A, editors. Biological evolution and statistical physics. Berlin: Springer. p 148-161.

9 Pagel MD, Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167:808-825.

10 Ives AR, Midford PE, Garland TJ. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56:252-270.

11 Hansen TF, Bartoszek K. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. Syst Biol 61:413-425.

12 Garamszegi LZ. 2014. Uncertainties due to within-species variation on comparative studies: measurement errors and statistical weights. In: Garamszegi LZ, editor. Modern phylogenetic comparative methods and their application in evolutionary biology - concepts and practice. Berlin: Springer. p 157-199.

13 Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341-1351.

14 Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683-695.

15 Borries C, Gordon AD, Koenig A. 2013. Beware of primate life history data: a plea for data standards and a repository. PLoS One 8:e67200.

16 Ioannidis JPA. 2005. Why most published research findings are false. PLoS Med 2:e124.

17 Freedman LP, Cockburn IM, Simcoe TS. 2015. The economics of reproducibility in preclinical research. PLoS Bio 13:e1002165.

18 Arnold C, Matthews LJ, Nunn CL. 2010. The 10kTrees website: a new online resource for primate phylogeny. Evol Anthropol 19:114-118.

19 Drummond CG. 2009. Replicability is not reproducibility: nor is it good science.

20 Begley CG, Ellis LM. 2012. Raise standards for preclinical cancer research. Nature 483:531-533.

21 Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspect Psychol Sci 7:652-655.

22 Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes E, Carbone C and others. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648.

23 Harvey PH, Martin RD, Clutton-Brock TH. 1987. Life histories in comparative perspective. In: Smuts BB, Cheney DL, Seyfarth RM, Wrangham RW, Struhsaker TT, editors. Primate societies. Chicago: University of Chicago Press. p 181-196.

24 Kappeler PM, Pereira ME, editors. 2003. Primate life histories and socioecology. Chicago: University of Chicago Press. 395 p.

25 Jackson G, Mooers AO, Dubman E, Hutchen J, Collard M. 2014. Basal metabolic rate and maternal energetic investment durations in mammals. BMC Evol Biol 14:194.

26 Bielby J, Mace GM, Bininda-Emonds ORP, Cardillo M, Gittleman JL, Jones KE, Orme CDL, Purvis A. 2007. The fast-slow continuum in mammalian life history: an empirical reevaluation. Am Nat 169:748-757.

27 Schultz AH. 1942. Growth and development of the proboscis monkey. Bull Mus Comp Zool 89:279-314.

28 Gordon AD. 2006. Scaling of size and dimorphism in primates II: Macroevolution. Int J Primatol 27:63-105.

29 Simoes TR, Caldwell MW, Palci A, Nydam RL. early view. Giant taxon-character matrices: quality of character constructions remains critical regardless of size. Cladistics.

30 Wheatley BP. 1980. Feeding and ranging of East Bornean *Macaca fascicularis*. In: Lindburg DG, editor. The macaques: studies in ecology, behavior and evolution. New York: van Nostrand Reinhold. p 215-246.

31 Tsuji Y, Hanya G, Grueter CC. 2013. Feeding strategies of primates in temperate and alpine forests: comparison of Asian macaques and colobines. Primates 54:201-215.

32 Jablonski NG, Whitfort MJ, Roberts-Smith N, Xu Q. 2000. The influence of life history and diet on the distribution of catarrhine primates during the Pleistocene in eastern Asia. J Hum

Evol 39:131-157.

33 Patterson SK, Sandel AA, Miller JA, Mitani JC. 2014. Data quality and the comparative method: the case of primate group size. Int J Primatol 35:990-1003.

34 Hassel-Finnegan HM, Borries C, Larney E, Umponjan M, Koenig A. 2008. How reliable are density estimates for diurnal primates? Int J Primatol 29:1175-1187.

35 Martin P, Bateson P. 2007. Measuring behaviour: an introductory guide. Cambridge: Cambridge University Press. 176 p.

36 National Academy of Sciences, National Academy of Engineering, Institute of Medicine of the National Academies. 2009. On being a scientist. A guide to responsible conduct in research. Washington DC: National Academies Press.

37 Martin RD, Genoud M, Hemelrijk CK. 2005. Problems of allometric scaling analysis: examples from mammalian reproductive biology. J Exp Biol 208:1731-1747.

38 Deaner RO, Nunn CL, van Schaik CP. 2000. Comparative tests of primate cognition: different scaling methods produce different results. Brain Behav Evol 55:44-52.

39 Smith RJ, Jungers WL. 1997. Body mass in comparative primatology. J Hum Evol 32:523-559.

40 Gordon AD, Marcus E, Wood BA. 2013. Great ape skeletal collections: Making the most of scarce and irreplaceable resources in the digital age. Yearb Phys Anthropol 152:2-32.

41 Borries C, Lu A, Ossi-Lupo K, Larney E, Koenig A. 2014. The meaning of weaning in wild Phayre's leaf monkeys: last nipple contact, survival, and independence. Am J Phys Anthropol 154:291-301.

42 Lee PC. 1996. The meanings of weaning: growth, lactation, and life history. Evol Anthropol 5:87-96.

43 van Noordwijk MA, Willems EP, Utami-Atmoko SS, Kuzawa C, van Schaik CP. 2013. Multi-year lactation and its consequences in Bornean orangutans (*Pongo pygmaeus wurmbii*). Behav Ecol Sociobiol 67:805-814.

44 Harvey PH, Clutton-Brock TH. 1985. Life history variation in primates. Evolution 39:559-581.

45 Dubman E, Collard M, Mooers AO. 2012. Evidence that gestation duration and lactation duration are coupled traits in primates. Biol Lett 8:998-1001.

46 Asquith PJ. 1989. Provisioning and the study of free-ranging primates: history, effects, and prospects. Yearb Phys Anthropol 32:129-158.

47 Whitlock MC. 2011. Data archiving in ecology and evolution: best practices. Trends Ecol Evol 26:61-65.

48 Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G and others. 2015. Promoting an open research culture - Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science 348:1422-1425.

49 Parker TH, Nakagawa S, Gurevitch J, IIEE. 2016. Promoting transparency in evolutionary biology and ecology. Ecol Lett 19:726-728.

50 Strier KB, Altmann J, Brockman DK, Bronikowski AM, Cords M, Fedigan LM, Lapp H, Liu X, Morris WF, Pusey AE and others. 2010. The primate life history database: a unique shared ecological data resource. Methods Ecol Evol 1:199-211.

51 Redmond E, Wilson JR. 2012. Seven databases in seven weeks: a guide to modern databases and the NoSQL movement. Dallas: Pragmatic Bookshelf.

52 Wilson G, Aruliah DA, Brown CT, Chung Hong NP, Davis M, Guy RT, Haddock SHD, Huff

KD, Mitchell IM, Plumbley MD and others. 2014. Best practices for scientific computing. PLoS Bio 12:e1001745.

53 Wieczorek J, Bloom D, Guralnick RP, Blum S, Doering M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin core: an evolving community-developed biodiversity data standard. PLoS One 7:e29715.

54 Sadleir RMFS. 1969. The ecology of reproduction in wild and domestic mammals. London: Methuen. 321 p.

55 Bronson FH. 1989. Mammalian reproductive biology. Chicago: University of Chicago Press. 325 p.

56 Leigh SR. 1994. Relations between captive and noncaptive weights in anthropoid primates. Zoo Biol 13:21-43.

57 Rafferty AR, Wong BBM, Chapple DG. 2015. An increasing citation black hole in ecology and evolution. Ecol Evol 5:196-199.

58 Takahata Y, Suzuki S, Agetsuma N, Okayasu N, Sugiura H, Takahashi H, Yamagiwa J, Izawa K, Furuichi T, Hill DA and others. 1998. Reproduction of wild Japanese macaque females at Yakushima and Kinkazan Islands: a preliminary report. Primates 39:339-349.

59 Altmann J, Alberts SC. 2005. Growth rates in a wild primate population: ecological influences and maternal effects. Behav Ecol Sociobiol 57:490-501.

60 Borries C, Koenig A, Winkler P. 2001. Variation of life history traits and mating patterns in female langur monkeys (*Semnopithecus entellus*). Behav Ecol Sociobiol 50:391-402.

61 Sandel AA, Miller JA, Mitani JC, Nunn CL, Patterson SK, Garamszegi LZ. 2016. Assessing sources of error in comparative analyses of primate behavior: Intraspecific variation in group size and the social brain hypothesis. J Hum Evol 94:126-133.